Bayesian methods for Model Selection and related problems

Gonzalo García-Donato

Universidad de Castilla-La Mancha (Spain) and VaBar

ScoVa16-Valencia

VaBar and model selection

- Model Selection, MS, has been a main line of research of our group during the last 15 years.
- Our work is greatly influenced by two prominent Bayesians: Susie Bayarri and Jim Berger.
- Our paradigm is objective Bayesian.
- Regarding the relation MS-VaBar, this talk presents the problem, reviews what we have done (**past**), briefly introduce current lines of research (**present**).
- The objective is call your attention to this fascinating problem and, why not, capturing your interest to collaborate in the next years.... simBIOSSis and let us write the **future!**

Model selection

- *Model Selection (or model choice), a definition*: statistical problem where several statistical models

$$M_1(\boldsymbol{y} \mid \boldsymbol{\theta}_1),\ M_2(\boldsymbol{y} \mid \boldsymbol{\theta}_2), \ldots, M_k(\boldsymbol{y} \mid \boldsymbol{\theta}_k),$$

are considered as plausible explanations for an experiment with output $\boldsymbol{y}$.

Keyword is *model uncertainty*

since it is unknown which is the true model and that uncertainty is explicitly considered.

- The set of competing models is called the model space and usually is denoted as $\mathcal{M}$.
- Possible specific MS goals:
  - To choose a single model (which is the true model?),
  - To explicitly incorporate model uncertainty to provide more realistic inferences/predictions (this is called *Model Averaging*), a problem sometimes presented as mixture modeling.
- Two particular (and very popular) MS problems
  - Hypothesis testing, and
  - Variable selection

Hypothesis testing

• In testing, the competing models have a common statistical form, say $M(\boldsymbol{y} \mid \boldsymbol{\theta})$, but differ on where $\boldsymbol{\theta}$ is located

$$M_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i) = \{M(\boldsymbol{y} \mid \boldsymbol{\theta} = \boldsymbol{\theta}_i),\ \boldsymbol{\theta}_i \in \Theta_i\},\ i = 1, \ldots, k,$$

normally denoted as

$$H_i : \boldsymbol{\theta} \in \Theta_i,\ i = 1, \ldots, k.$$

• Particularly popular/important/difficult is the testing problem with some of Θ_i consisting on a single point in the corresponding Euclidean space (e.g. $\theta = 0$). This testing problem is normally called *point* or *precise*.

Variable selection

Variable selection

- Model selection problems where the different models, M_i differ about which variables of a given set $x_1, x_2, \ldots, x_p$ explains a response variable y.

Variable selection is a multiple testing problem with 2^p (precise) hypotheses of the type

$$H_i : \beta_{j_1} = \cdots = \beta_{j_k} = 0.$$

Posterior model probabilities

• The formal Bayesian answer to the MS problem is based on the posterior probabilities of the competing models:

$$Pr(M_j \mid \boldsymbol{y})$$

• Such (discrete) posterior distribution encapsulates the responses to every question in Model Selection. Two examples:

- If you want to select a single model use the most probable a posteriori and report its posterior probability as a measure of uncertainty.
- Model averaging? Use $Pr(M_j \mid \boldsymbol{y})$ as weights.

Posterior model probabilities and Bayes factors

Assuming one of the models in $\mathcal{M}$ is the true model

$$Pr(M_j \mid \boldsymbol{y}) = \frac{m_j(\boldsymbol{y})Pr(M_j)}{\sum_i m_i(\boldsymbol{y})Pr(M_i)} = \frac{B_{j0}Pr(M_j)}{\sum_i B_{i0}Pr(M_i)},$$

where

Ingredient	Name	Type of problem
$m_i(\boldsymbol{y}) = \int M_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\, \pi_i(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$	prior marginal	Computational
B_{i0}	Bayes factor of M_i to M_0	None
$Pr(M_i)$	prior prob. of M_i	Multiplicity
C	Normalizing constant	Computational
$\pi_i(\boldsymbol{\theta}_i)$	prior for M_i	All sort of problems

$\pi_i(\boldsymbol{\theta}_i)$: the main conceptual challenge

MS prior distributions are, perhaps, the most problematic aspect of any Bayesian approach. In MS the difficulties grow:

- Results are very sensitive to the prior distribution (changing the prior you can essentially obtain whatever you want).
- Such sensitiveness does not disappear asymptotically with n.
- Neither improper nor vague priors can be used.
- Frequentist properties are not very useful to differentiate among priors (and are potentially misleading)

Contributions (I): About $\pi_i(\boldsymbol{\theta}_i)$

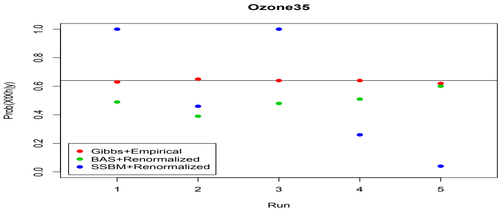
Our contributions about π_i have roots on Jeffreys who first proposed using proper priors centered at the 'null' and with flat tails (Cauchy). Zellner and Siow (1980) extended this idea to regression problems.

- Bayarri and Garcia-Donato (2007), used such priors to test general hypotheses in linear models (regression or ANOVA).
- Bayarri and Garcia-Donato (2008), propose a general mathematical rule that extend Jeffreys' priors to any other testing problem. These are named as Divergence Based priors.
- Bayarri, Berger, Forte and Garcia-Donato (2012), introduce a deep methodological change. They propose and formalize the idea of specifying (and characterizing) priors based on sensible criteria like invariance, predictive matching and consistency.
- The method is illustrated proposing a new prior for variable selection in linear models with optimal properties that they call *Robust prior*.

Contributions (II): Computational aspects

The number of competing models in variable selection, 2^p, becomes easily very, very large. Posterior probabilities cannot be exactly computed and heuristic methods are called for.

- Garcia-Donato and Martinez-Beneito (2013), show that very simple Gibbs algorithms plus frequency of visits to estimate probabilities, largely outperforms modern searching methods with estimations based on re-normalization.



Contributions (III): Software

The high specificity of the MS problem jointly with the particularities of its priors makes almost useless standard Bayesian software (of the type WinBUGS, etc)

• Forte and Garcia-Donato (2012) have developed `BayesVarSel`, an `R`-package that solves testing and variable selection problems in linear models.
Main characteristics:

- Priors: "Robust", "g-Zellner", "Zellner-Siow", "Liang".
- Priors for M_i: "Constant", "ScottBerger".
- Methods: exact (sequential or parallel) and the empirical Gibbs here cited.
- Results: HPM, inclusion probabilities (univariate, joint, conditionals), image plots, etc.
- And with a simple and familiar (`lm`-type) interface:
```
>Bvs(formula="IMC~ .", data=obesity, n.keep=1000)
```

Contributions (IV): Applications

Does it have real applications?

- Determining the number of jointpoints in epidemiological temporal series (Martinez-Beneito and others in 2012).
- Studying which factors explain the Gross Domestic Product in the US (Forte and others in 2015).

1. Introducing the problem
2. Our view and what we have done
3. What we are doing now
   - Large p small n problem
   - Other projects in progress

1. Introducing the problem
2. Our view and what we have done
3. What we are doing now
   - Large p small n problem
   - Other projects in progress

High dimensional setting

• With M. Martinez-Beneito and in collaboration with Jim Berger (Duke University) we are working on the variable selection problem with $n << p$.
• Key idea is noting that models can be classified as singular (more parameters than n) or regular (number of parameters $\leq n$). Hence $\mathcal{M} = \mathcal{M}^{\mathcal{S}} \cup \mathcal{M}^{\mathcal{R}}$.

Result

the Bayes factor of any $M_\gamma \in \mathcal{M}^S$ to the null is $B_{\gamma 0} = 1$ (for the rest $B_{\gamma 0}$ is the conventional one, say robust).

$\mathcal{M}^S$ (dark side)

$\mathcal{M}^R$ (light side)

Who wins?

The scene now is that the dark side, $\mathcal{M}^S$, is a vast deserted (only ones) region while the light side, $\mathcal{M}^R$, is a minuscule region lighted by the data.

• *eg. if $p = 8408$ and $n = 41$, the proportion of regular models over the total number of models is of the order 10^{-2000}.*

The relevance a posteriori of singular models is

$$P^S = Pr(M^T \in \mathcal{M}^S \mid \boldsymbol{y}) = \frac{p-n+1}{p-n+1+n\mathcal{C}^R},$$

where $\mathcal{C}^R$ is the normalizing constant conditionally on $\mathcal{M}^R$.

The methodology in practice

No need to ‘explore’ the whole model space, it suffices with

- exploring $\mathcal{M}^R$ (still moderate to large: MCMC in Garcia-Donato and Martinez-Beneito, 2013),
- estimating $\mathcal{C}^R$ (George and McCulloch, 1997) and hence P^S,
- any relevant feature of the posterior distribution can be easily computed.

An illustrative example

Simulated experiment in Hans et al (2007), with $n = 41$ patients and $p = 8408$ genes from a tumor specimen. The 'true' data generating model is

$$y_i = 1.3x_{i1} + .3x_{i2} - 1.2x_{i3} - .5x_{i4} + N(0, 0.5).$$

We obtain:

n	P^S	q_1	q_2	q_3	q_4	$\bar{q}_{-T}$	q^U_{-T}	HPM
41	0.004	0.843	0.154	0.766	0.002	0.002	0.038	$\{x_1, x_3\}$
30	0.320	0.300	0.171	0.173	0.160	0.159	0.251	$\{x_1, x_3\}$
20	0.830	0.417	0.416	0.415	0.415	0.414	0.419	$\{x_{4026}, x_{7748}\}$
10	0.995	0.497	0.497	0.497	0.497	0.497	0.497	{Null,Full}

Keys: q_i is the inclusion probability for x_i ($i = 1, 2, 3, 4$) and $\bar{q}_{-T}$, q^U_{-T} are respectively the mean and maximum of the inclusion probabilities for the spurious variables. HPM is the estimated most probable a posteriori model.

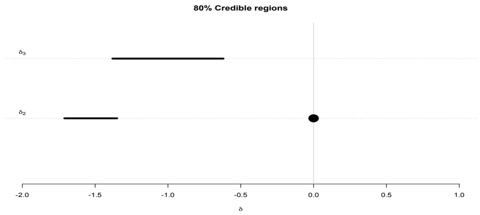
BayesVarSel vs. others

- There are other R packages, that perform similar calculations as ours does.
- The closest are BayesFactor, BMS and mombf.
- With A. Forte and in collaboration with Mark Steel (Warwick) we are working on comparing these packages with emphasis on compatibility and efficiency.

Package	BayesFactor	BayesVarSel	BMS
	newPriorOdds(BFobject)=	prior.models=	mprior=
$\theta = 1/2$	rep(1,2^ p)	"constant"	"fixed" or "uniform"
$\theta \sim$ Unif(0,1)	-	"ScottBerger"	"random"

Variable selection in ANCOVA

- Analysis of Covariance (ANCOVA) models are linear models with continuous (also known as covariates) and categorical (factors) explanatory variables.
- A factor F with J levels enters the design matrix through $x_1, \ldots, x_{J-1}$ dummy variables. This implies several problems for the variable selection strategy.
  - Prior probabilities?
  - How to interpret inclusion probabilities? Which level is causing the factor be significant?
  - How to summarize results?
- This is a line of research in collaboration with R Paulo (U of Lisbon)



Students

- With A Forte and A Moro (master thesis work): The theory of the criteria paper, (Bayarri et al, 2012), although presented generically, was illustrated in normal Linear models. How does it apply in GLM's?
- With A Forte and E Moreno (thesis in progress): the problem with missing data in variable selection and summarizing the posterior distribution.
- With A Forte and R Gavidia (thesis in progress): Variable selection in genomics.

VaBar and model selection

Thanks!